Egy nyelvészeti UIMA-folyamat a kézi annotálástól az eredmények megjelenítéséig

Kiss Márton[1], Nagy Ágoston[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
H-6720 Szeged, Árpád tér 2.
{mkiss, nagyagoston}@inf.u-szeged.hu

Kivonat: A MaSzeKer projekt indulásakor az UIMA nyelvészeti keretrendszert választottuk a fejlesztéshez. Az már a fejlesztés kezdetekor látszott, hogy a következő modulokra mindenképpen szükségünk lesz a projekt nyelvészeti részének megvalósításához: *kézi annotálás, gépi annotálás, a két annotáció összehasonlítása* és az *eredmények megjelenítése.* Ezen igények teljes körű kielégítésére nem találtunk implementált rendszert. Ezért kifejlesztettünk egy nyelvészeti UIMA-folyamatot támogató környezetet (UIMA-modulokat és hozzájuk kapcsolódó segédprogramokat), mely az előbb említett technikai elvárásokat megvalósítja. A cikkben bemutatjuk a létrejött rendszer mindazon részeit, melyek segítségével nyomon követhető, segíthető egy nyelvészeti kutatás a dokumentumok kézi annotálásától az eredmények megjelenítéséig.

1 Bevezetés

Cikkünkben bemutatjuk azokat a kifejlesztett UIMA-modulokat és segédprogramokat, melyek segítségével megvalósítottunk egy olyan rendszert, mely hatékony támogatást nyújt számítógépes nyelvészeti kutatásokhoz. A kifejlesztett rendszernek azon részeit mutatjuk be, melyek támogatják: kisebb méretű, pár száz dokumentumot tartalmazó *tanulókorpusz építését,* a korpuszon végzett gépi és kézi jelölések *összehasonlítását*, valamint az eredmények *vizualizált megjelenítését*. A kifejlesztett rendszer segítségével könnyen megtalálhatjuk a gépi rendszer hiányosságait és kijavíthatjuk az esetleges hiányosságokat.

2 Az UIMA-modulok és a segédprogramok bemutatása

A következő fejezetben végigvesszük a modulok és segédprogramokat, ismertetve a működésüket és technikai megvalósításukat. A fejezet végén pedig egy ábrán szemléltetjük a rendszer felépítését és a modulok kapcsolatát.

A kifejlesztett UIMA-modulok: *AnnotationComparator, HTMLViewer*. A segédprogramok: *Word <-> UIMA konverter, Word <-> TXT konverter*.

2.1 Kézzel annotált Word-dokumentum konvertálása UIMA XMI-be (Word-makró + Perl + UIMA-modul)

A nyelvész kollégák számára olyan annotálási módszert kellett kidolgoznunk, mely könnyen elsajátítható és kényelmesen végezhető vele a munka. Erre azt a megoldást találtuk a legalkalmasabbnak, hogy egy Word-dokumentumban jelöljék meg a releváns szövegrészeket valamilyen előre megállapított formázás segítségével, például változtassák meg a szöveg háttérszínét. Ezek után a Word-dokumentumból az UIMA számára is értelmezhető annotációkat kellett készíteni. Ezen technikai megoldással létrejöttek a tanulókorpuszok.

A modul első lépésben kiexportálja a Word-dokumentumokban formázással jelölt kézi annotációkat egy *egyszerű XML*-fájlba (Word -makró segítségével). Ezek után az XML-fájlból egy Perl-script segítségével olyan *konfigurációs fájlokat* készítünk, melyek tartalmazzák az annotációkat és a pontos karakterpozícióját a szövegben. Végül egy UIMA-modul a konfigurációs fájl segítségével létrehozza az *annotációkat.*

2.2 Word-dokumentum, TXT-konverter (Word-makró)

A Word-dokumentumot TXT formátumra is kellett hozni, hogy az UIMA rendszer moduljai bemenetként felhasználhassák. Ezt a problémát egy Word-makró segítségével oldottuk meg, mely egy könyvtár (egy korpusz) összes .doc kiterjesztésű fájlját átalakítja TXT formátumra.

2.3 Annotációk összehasonlítása (UIMA-modul)

Amikor előálltak a kézi és gépi annotációk is, szükségünk volt arra, hogy összehasonlítsuk a kettőt. A gépi algoritmus hatékonyságát a pontosság, a fedés és az F-mérték kiszámításával mértük. Az összehasonlítás során többféle illeszkedés is beállítható attól függően, hogy hogyan szeretnénk összehasonlítani a két annotációt. Választható illeszkedési típusok:

teljes: ekkor a két annotációnak teljesen meg kell egyeznie mind a kezdeti, mind a végső karakterpozícióban

tartalmaz: ebben az esetben a két annotáció akkor is egyezik, ha az egyik „csak” tartalmazza a másikat, vagyis annot1 és annot2 annotációk esetén: annot1.begin <= annot2.begin és annot2.end <= annot1.end

Az összehasonlítás során a további hatékonyság növelése érdekében az összehasonlító modul kigyűjti a rosszul bejelölt vagy nem megjelölt annotációkat.

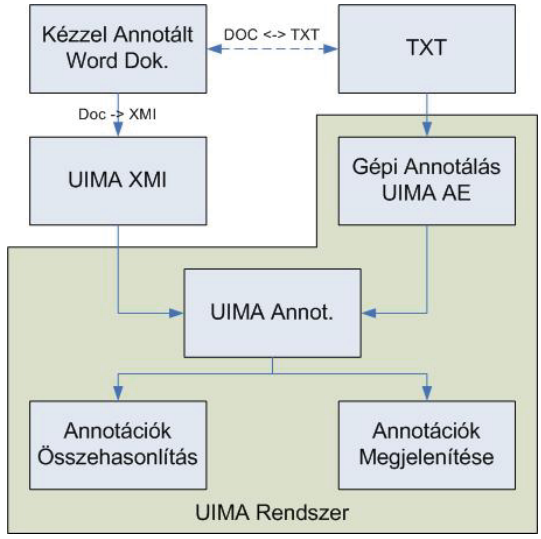
2.4 Megjelenítés (UIMA-modul + Perl + JavaScript)

Az eredmények vizualizációjára az ellenőrzés és az átláthatóság miatt volt szükség. Kétféle megjelenítőt készítettünk:

a) AZ UIMA InLine XML megjelenítése (XSL)

b) AZ UIMA XMI megjelenítése (Perl+UIMA+HTML)

Az a) esetben azon adatok, annotációk, kapcsolatok megjelenítésére van lehetőség, melyek *fastruktúrában* ábrázolhatóak: szülő és ős kapcsolat áll fenn két annotáció között. A b) esetben a feldolgozás során *bejelölt összes annotáció* megjeleníthető HTML formátumban.



1. ábra. A megvalósított UIMA-modulok és a segédprogramok kapcsolata.

Bibliográfia

1. Kano, Y., Nguyen, N., Sætre, R., Yoshida, K., Miyao, Y., Tsuruoka, Y., Matsubayashi, Y., Ananiadou, S., Tsujii, J.: Filling the gaps between tools and users: a tool comparator, using protein-protein interaction as an example. In: Proceedings of Pacific Symposium on Biocomputing (PSB), 13 (2008) 616–627
2. Ferrucci, D., Lally ,A.: Building an example application with the Unstructured Information Management Architecture. IBM Systems Journal Vol. 43 No. 3 (2004) 455–475
3. Kano, Y. et al.: U-Compare: share and compare text mining tools with UIMA. Bioinformatics, doi: 10.1093/bioinformatics/btp289 (2009)
4. D. Ferrucci, A. Lally: UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Journal of Natural Language Engineering Vol. 10 No. 3-4 (2004) 327–348
5. Kunze, M., Rösner, D.: Tools for UIMA Teaching and Development. University of Magdeburg, Germany (2008)